基于聚类的星系光谱分析*

张茜[1,2]，张健楠[1]，赵永恒[1]
(1.中国科学院光学天文重点实验室（国家天文台），北京 100101;
2. 中国科学院大学，北京 100049)

摘 要：各类大型巡天项目产生了海量天文数据，因此需要研究适用于大规模数据的光谱自动处理方法。传统的基于谱线检测或 BPT 图的星系光谱分类方法难以直接应用于星系光谱自动分类 pipeline，相比之下基于机器学习的光谱自动分析更适用于海量天文数据的分类研究。本文提出了一种基于双层聚类的星系光谱分析方法。第一层采用 k 均值聚类算法将星系光谱聚为吸收线星系和发射线星系，第二层使用 CLARA（Clustering LARge Applications）聚类算法将发射线星系聚为 5 簇。对 LAMOST DR5 的星系数据进行实验，其结果表明：（1）第一层 k 均值聚类能够成功将星系光谱分为吸收线星系和发射线星系，聚类簇与基于谱线检测的分类结果基本一致。（2）第二层 CLARA 聚类结果能够在 BPT 图中反映出不同的星系类型。（3）光谱聚类结果与颜色星等图分类存在预期的相关性。（4）k 均值聚类和 CLARA 聚类能够适用于大规模数据自动分析处理，聚类结果能够很好地反映出星系的物理性质和演化过程，簇心数据可以为光谱自动分类 pipeline 提供模板。

关键词：LAMOST； 聚类； 星系光谱分类；大样本光谱分析；
中图分类号：P157.1 **文献标识码：**A **文章编号：**

0 引 言

星系光谱分类对于研究星系的形成与演化具有重要意义。传统星系分类方法包括：基于形态学的哈勃分类法，根据星系外形将星系分为椭圆星系、旋涡星系、棒旋星系和不规则星系；基于颜色的分类法，Strateva[1]分析 SDSS 数据时发现颜色星等图服从双峰分布，蓝色星系和红色星系各有峰值，双峰之间为绿谷；以及基于光谱的 Baldwin，Phillips，Terjevich（BPT）诊断图[2]的分类方法，经过多年的改进形成了基于线强比诊断图的分类方法，目前常用经验分割线有 Kauffmann 提出的用于识别纯恒星形成星系（Star-Forming，简称 SF）的分割线[3]，Kewley 等人提出的用于识别纯活动星系核星系（AGN）的分割线[4]，以及 Kewley[5]和 Cid Fernandes[6]分别提出的用于区分 LINER（Low-Ionization Nuclear Emission-Line Region）星系和 Seyfert2 星系的分割线。

大型巡天项目的实施为天文领域提供了海量光谱数据，例如 2dF、6dF、RAVE、SDSS、LAMOST、GAIA 等，其中 LAMOST DR5 发布星系光谱多达 15 万余条，必须研究光谱自动分类技术用于大规模光谱数据的分类研究。传统的基于谱线检测或 BPT 图的星系光谱分类方法需要进行星族成分合成，由于此过程复杂且耗时，不适用于海量光谱数据的处理，无法直接用于光谱自动分类 pipeline，相比之下，基于机器学习的光谱自动分类方法更适用于海量天文数据的分析研究。目前有许多机器学习方法成功应用于天体分类的案例，包括监督型和无监督型分类方法。无监督型的分类方法有主成分分析（PCA）法，它广泛运用于星系光谱的识别与分类中，例如 SLOAN 巡天项目中的光谱处理系统就是利用星系光谱主成分进行星系光谱的识别[7]，另外 Almeida[8]成功将 k 均值方法应用于星系光谱分类中，分类结果

* **基金项目：**国家自然科学基金（11403059）；国家自然科学基金天文联合基金（U1531242）资助.
收稿日期：年-月-日；**修订日期：**年-月-日：
作者简介：张茜，女，硕士生，研究方向：天体光谱的自动分析处理。 Email:zhangxi@bao.ac.cn
通讯作者：张健楠，女，副研究员，研究方向：天文数据处理。 Email:jnzhang@bao.ac.cn

能很好地体现星系演化过程。监督型分类方法有许多，例如文[9]使用基于 Fisher 判别分析的有监督特征提取方法对类星体和正常星系分类，文[10]使用支持向量机方法对活动天体和非活动天体分类，文[11]使用决策树方法对星系形态学分类。

聚类属于无监督型方法，具有算法简单、收敛速度快和准确率高的特点。聚类主要依赖于数据特征进行自动分类，过程独立且受主观因素影响小，相较于监督型方法，不需要提供已有标签数据进行训练，同时聚类结果中数量较少的簇有助于发现稀有天体。本文针对 LAMOST DR5 中星系光谱数据，设计了双层聚类方法对星系光谱进行聚类分析。本文结构如下：第 1 节为双层聚类方法介绍，第 2 节介绍了星系光谱聚类实验，包括预处理方法、实验步骤和参数选择等，第 3 节对实验结果进行分析，从聚类结果是否有效和是否具有物理性质两方面分析，将聚类实验结果与基于谱线检测、BPT 图和颜色星等图的分类结果进行比较，第 4 节为结论。

1 双层聚类方法

针对星系光谱特点和不同聚类算法的特点，本文提出了双层聚类方法对星系光谱进行聚类分析。第一层采用 k 均值聚类算法[12]将星系光谱分为吸收线星系和发射线星系，k 均值聚类算法简单，能够快速收敛，对于大数据处理具有伸缩性，适用于大规模星系光谱处理。第二层采用 CLARA 聚类算法[13]将发射线星系聚为 5 个子类，CLARA 算法简单，对噪声不敏感，适用于大规模数据。

1.1 k 均值聚类算法

k 均值（k-means）聚类算法的核心内容就是将数量为 n 的样本划分为 k 类，并且每个样本点到聚类中心的距离平方和最小。

k-means 算法基本步骤如下：

输入：n 个样本和聚类个数 k。

输出：将样本划分为 k 类。

（1）从 n 个样本中选取 k 个初始点作为初始聚类中心；

（2）计算每个样本点与聚类中心的距离，将样本划分到距离它最近的聚类中心所属的类；

（3）重新计算每一类中所有样本点的平均值作为新的聚类中心，并计算每个样本点到它所在类的聚类中心的距离平方和 D；

（4）判断聚类中心和 D 是否改变，若改变，更新聚类中心后重复 2、3 步，否则聚类结束。

影响聚类效果的因素有很多，k 值的选取、初始聚类中心的选取方法以及距离测度方法都会影响聚类效果。k 值的选取方法包括凭经验选取和按密度选取。挑选初始聚类中心常用的方法有四种。一是随机选取 k 个样本作为初始聚类中心；二是随机采用样本空间中 10% 的数据做预聚类，预聚类的初始聚类中心也是随机挑选的；三是根据样本的取值范围均匀的随机选取 k 个聚类中心；四是考虑权重的 k-means++方法，随机选取第一个聚类中心后，计算所有点到此聚类中心的距离，将距离作为权重来选择下一个聚类中心，目的是使距离大的点被选中的概率更大一些，然后重复选取 k 个聚类中心。距离度量方法有：欧氏距离、曼哈顿距离、余弦距离和相关距离等。

本文聚类实验中，在考虑到光谱的特点并对比多种距离后选取相关距离作为距离度量方法，相关距离为 $d = 1 - \rho$，其中 ρ 为相关系数，用于判断随机变量 X 与 Y 的相关程度，其表达式为：

$$\rho = \frac{\operatorname{cov}(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} = \frac{E((X-E(X))(Y-E(Y)))}{\sqrt{D(X)}\sqrt{D(Y)}} \quad (1)$$

ρ 取值范围为[-1,1]，绝对值越大，表明 X 与 Y 的相关度越高。

1.2 CLARA 聚类算法

K-means 聚类算法对噪声敏感度高，k 中心点（k-medoids）[14]聚类是对 k-means 的改进，k-means 算法更新聚类中心是求取类内平均值，而 k-medoids 将每个点代替聚类中心，降低离群点对聚类结果的影响。

k-medoids 算法基本步骤如下:

输入：*n* 个样本和聚类个数 *k*。

输出：将样本划分为 *k* 类。

（1）从 *n* 个样本中选取 *k* 个初始点作为初始聚类中心；

（2）计算所有样本点到聚类中心的距离，将样本划分到距离最近的聚类中心所在的类；

（3）随机选择一个非聚类中心点，计算此点代替原聚类中心的总代价，重复此步骤直到所有非聚类中心点都被判断过；

（4）判断每个非聚类中心点代替原中心点的总代价，若有小于 0 的，从中挑选出总代价最小的一个所对应的非聚类中心点，将此点作为新的聚类中心；

（5）重复（3）、（4）步骤，直到聚类中心点不变，聚类结束。

判断能否用新的非聚类中心点 *Oh* 代替原聚类中心点 *Oi*，对于每一个非中心点 *Oj* 都要满足如下规则：无论 *Oj* 原来属于 *Oi* 类还是另一个 *Om* 类，当 *Oh* 替换 *Oi* 后，*Oj* 会分配给距离它最近的类，可以是 *Oi* 或 *Om*，也可以是新的类 *Oh*。

新的非聚类中心点 *Oh* 代替原聚类中心点 *Oi* 的总代价是所有非中心点对象产生的代价之和。计算公式如下：

$$TC_{ih} = \sum_{j=1}^{n} C_{jih} \quad (2)$$

其中，*Cjih* 表示 *Oj* 在 *Oi* 被 *Oh* 代替后产生的代价，即 *Oj* 到原聚类中心的距离与 *Oj* 到新聚类中心的距离之差。若总代价为负，*Oi* 能被 *Oh* 替换，若总代价为正，则说明原聚类中心 *Oi* 不需要变化。

由于 k-medoids 聚类算法需要穷举类内点以达到寻找最优解的目的，此方法只适用于小规模数据。CLARA（Clustering Large Applications）是对 k-medoids 聚类算法的改进，用抽样样本代表全部数据计算聚类中心，能够应用于大规模数据聚类。

CLARA 算法基本步骤如下:

输入：*n* 个样本，聚类个数 *k*，抽样次数 *m*。

输出：将样本划分为 *k* 类。

（1）重复 *m* 次从全部样本中抽取（*40+2k*）个样本，每次重复执行（2）~（4）步骤；

（2）对此样本集使用 k-medoids 聚类，选出 *k* 个聚类中心；

（3）计算全部样本中每个非聚类中心点到聚类中心的距离，将其划分到距离最近的聚类中心所在的类；

（4）计算（3）步中的总代价，若小于当前值，则此聚类中心作为最佳聚类中心应用于全部样本，否则返回步骤（1），开始下一循环。

2 星系光谱聚类实验

2.1 数据预处理

本文采用的数据是从 LAMOST DR5 的 153093 条星系光谱中随机选取的 30000 条光谱。

因为缺少相应的测光设备，LAMOST 采用相对流量定标，即选择质量较好的 F 型矮星作为标准星，得到仪器的响应曲线，但是这些标准星的红化可能导致连续谱的不确定性，因此，需要对光谱进行重定标。本文采用 SLOAN 的 u,g,r,i,z 波段的 fiber 星等，在一定程度上校正 LAMOST 的连续谱。

重定标之后对光谱进行退红移处理，将其移至静止波长后，对光谱进行重采样，采样波长区间为 3600-9000Å，采样间隔为 1Å。

为避免噪声、环境等因素的影响，需要对光谱进行流量标准化，本文采用 S_{unit} 标准化方法。假设 x 是一条光谱，记为 $x=(x_1,x_2,\cdots,x_n)^T$，它是 n 维欧氏空间中的一个向量，流量标准化方法为[9]：

$$y = x / \sqrt{\sum_{i=1}^{n} x_i^2} \qquad (3)$$

在去除无法进行重定标和红移为坏值的光谱后，剩余 27272 条星系光谱用于聚类实验。

2.2 聚类实验

使用 k-means 聚类算法和 CLARA 聚类算法对 LAMOST DR5 中星系光谱进行聚类。实验分为两层，第一层用 k-means 将星系光谱分为吸收线星系和发射线星系，第二层用 CLARA 将发射线星系光谱细分类。

第一层，使用 k-means 聚类算法，将预处理后的 27272 条星系光谱分为发射线星系和吸收线星系。以年老恒星为主的早型星系的光谱以吸收线为主，发射线很弱甚至无法被探测到，相对年轻的晚型星系中有一部分与早型星系相似，发射线很弱，更晚型的星系中吸收线逐渐失去主导地位，发射线越来越明显。为使发射线和吸收线特征更为突出，将光谱去除连续谱。这里采用中值滤波方法拟合连续谱，用光谱流量减去连续谱得到谱线信息，对谱线信息进行聚类。

考虑到还有同时具有发射线和恒星成分的一类星系，选取 k 值为 3，用 k-means++方法获取初始聚类中心，使用相关距离作为距离度量方法。

第二层，使用 CLARA 聚类算法，将第一层聚类得到的发射线星系再进行细分类。连续谱可以反映出一部分发射线星系的特征，因此这一层聚类不需要去除连续谱。选取 r 波段信噪比大于 5 的共 12689 条星系光谱。为避免天光线的影响，用中值滤波法去噪，滤波窗口宽度为 5。考虑到一部分样本仅在波长为 3600-7900Å 有流量值，且 CLARA 聚类算法依赖于样本点，所以选择 3600-7900Å 范围内的光谱进行实验。

抽样次数为 100，使用相关距离作为距离度量方法。为选取较优的 k 值，画出 SSE（簇内误差平方和）随 k 值变化曲线，依据肘部法则，在 k=5 时观察到明显肘型，因此选取 k=5。

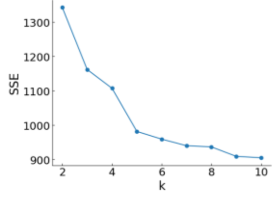


图 1 SSE 随 k 值变化图

Fig.1 The graph of SSE changing with k value.

3 星系光谱聚类结果分析

3.1 第一层聚类结果分析

k-means 聚类算法将 27272 条星系光谱分为三簇 cluster1，cluster2，cluster3，通过每一簇的聚类中心（图 2）可以看出其星系类型。发射线星系光谱以发射线为主，cluster1 发射线明显，为恒星成分很弱的强发射线星系，吸收线星系光谱吸收线占主导地位，发射线很弱甚至无法被探测到，由此看出 cluster2 属于吸收线星系，cluster3 发射线弱，为有恒星成分的弱发射线星系。

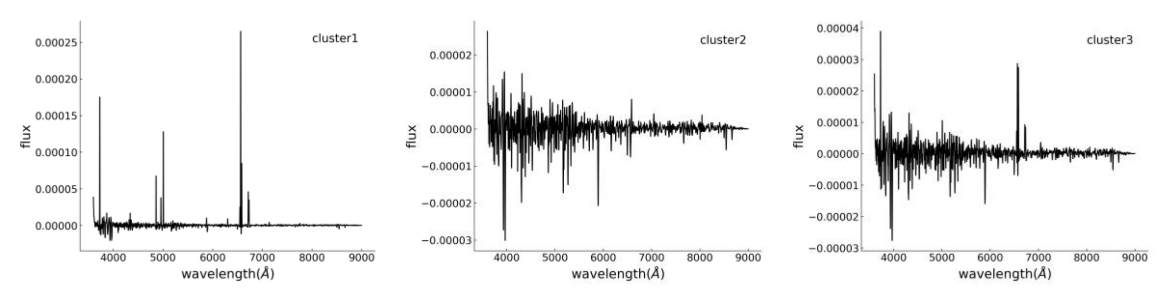


图 2 第一层聚类的聚类中心。左、中、右图分别为 cluster1、cluster2 和 cluster3 的聚类中心

Fig.2 The clustering centers of the first layer.The clustering center of cluster1,cluster2 and cluster3 are shown on the left,middle and right.

为探究聚类的稳定性，将 k-means 聚类方法应用于不同信噪比子集，分别从 27272 条星系光谱中取 r 波段信噪比大于 5、10、15、20 的四个子集，分别包含 23465、15593、9120、5166 条光谱数据。将 k-means 用于每个子集，得到的聚类中心见图 3，图 3 中四行图分别为 r 波段信噪比大于 5、10、15、20 的四个子集的聚类中心，为了便于比较将得到的聚类中心分别按发射线星系、吸收线星系和弱发射线星系排列，三列分别为 cluster1、cluster2 和 cluster3 簇的聚类中心，*mem* 表示此类所含样本个数，由不同子集的聚类中心都能反映出发射线星系、吸收线星系和弱发射线星系可以看出，k-means 聚类算法能够稳定聚类出这三种星系。

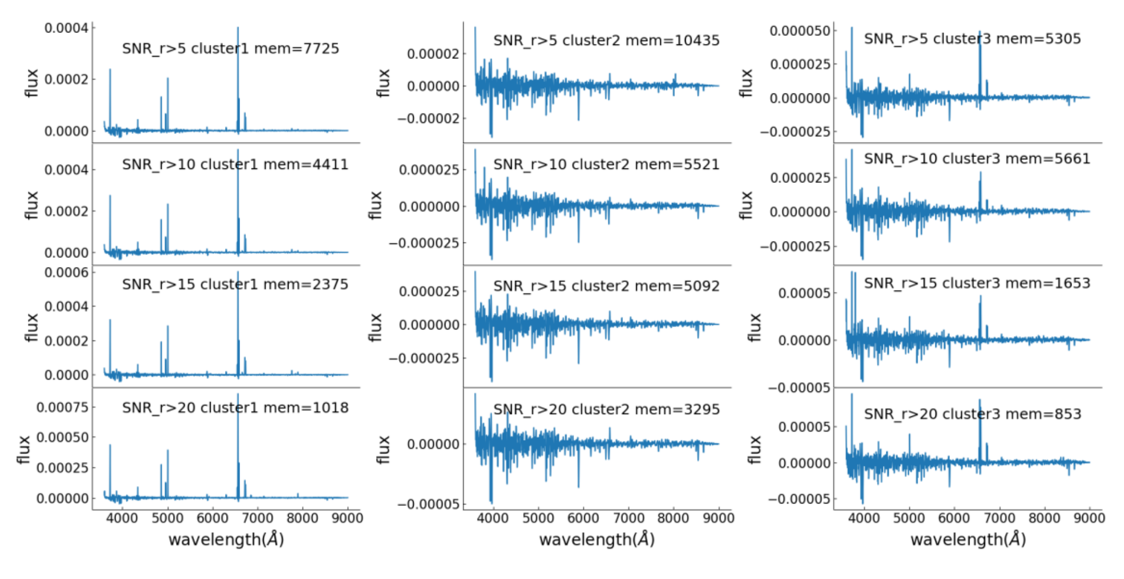


图 3 不同信噪比子集的聚类中心。四行由上至下分别为 r 波段信噪比大于 5、10、15、20 的四个子集的聚类中心，三列分别为每个子集的三个聚类中心，其中 mem 表示此类所含光谱数。

Fig.3 The clustering centers of different SNR subsets. The four rows from top to bottom are the cluster centers of the four subsets with r-band SNR greater than 5, 10, 15, and 20, and the three columns are the three cluster centers of each subset, where *mem* indicates the number of data in the cluster.

计算每一条光谱与每个聚类中心的距离，第 i 个簇 cluster i 的每一个样本与第 j 个聚类中心 center j 的距离统计图见图 4，其中三列图分别为三个簇中每一个样本与聚类中心的距离统计图，不同颜色代表不同信噪比数据集。整体来看，cluster i 与其本身的聚类中心距离相较于其他聚类中心更近。由图 4 中左列可以看出簇 cluster 1 与 center1 的距离靠近 0，与另两个聚类中心距离远，明显的三个峰表明第一个簇与另两个簇区分度明显。簇 cluster2 和 cluster3 在同一信噪比子集下，距离其本身的聚类中心距离更近，如图中第二列 cluster2 在信噪比大于 0 时（红色），距离 center1-3 的统计图峰值分别为 1、0.65、0.8。虽然簇 cluster2 和 cluster3 与其类内聚类中心的距离分布没有接近 0，但是从不同信噪比子集下的距离分布可以看出，随着信噪比的提高，簇 cluster2 和 cluster3 与其类内聚类中心的距离越来越靠近 0，如 cluster2-center2 图中，随着信噪比的提高，峰值从 0.65 降至 0.4。

每个样本与聚类中心相关距离分布也代表着类内距离分布，类内光谱的叠加得到的聚类中心信噪比提高，与相对信噪比较低的样本数据的相关性达不到 1，所以 cluster2-cneter2 和 cluster3-center3 的距离分布没有接近 0。从这个分布情况也可以看出 cluster2 和 cluster3 的类内分布不够紧致。

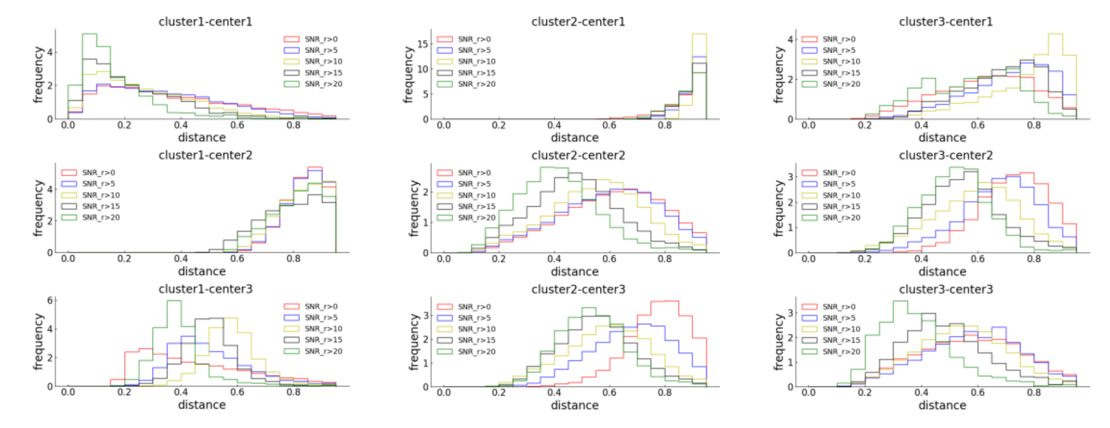


图 4 第一层聚类簇与聚类中心的距离统计图。图为第 i 个簇 cluster i 的每一个样本与第 j 个聚类中心 center j 的距离统计图，颜色表示不同信噪比数据集。

Fig.4 The distance statistical graph of the clusters and the cluster centers of the first layer. The figure shows the distance statistics of each sample of the i-th cluster cluster i and the j-th cluster center center j, and the colors represent different signal-to-noise ratio data sets.

将此聚类结果与传统分类方法的结果进行比较。传统区分吸收线星系和发射线星系，常使用 $S/N_\lambda \geq 3$ 作为判断依据，这里 S/N_λ 为谱线 λ 的信噪比。文[3-4]筛选发射线星系对 $H\alpha$、$H\beta$、$[OIII]\lambda5007$ 和 $[NII]\lambda6585$ 四条谱线都采用 $S/N_\lambda \geq 3$ 的筛选条件，但 Cid Fernandes[6] 等人发现，对四条谱线都进行筛选会使一些弱发射线星系被忽略，所以本文只对 $H\alpha$ 进行筛选。

聚类结果中 cluster1 和 cluster3 为发射线星系，cluster2 为吸收线星系，与用 $H\alpha$ 分类的结果进行比较（表 1），聚类结果与用 $H\alpha$ 分类的结果一致的数目在聚类每一类中的占比分别为：97.79%、80.80%、84.52%。对于全部数据，k-means 聚类结果中有 89.0%的星系与 $H\alpha$ 分类结果一致。

表 1 k-means 聚类结果与 $H\alpha$ 筛选结果数目比较

Tab.1 The comparison of the number between k-means and $H\alpha$ detection

数目（类内百分比）	$H\alpha$ 筛选为发射线星系	$H\alpha$ 筛选为吸收线星系	总计
cluster1（发射线星系）	12109（97.79%）	274（2.21%）	12383
cluster2（吸收线星系）	2169（19.20%）	9126（80.80%）	11295
cluster3（发射线星系）	3038（84.52%）	556（15.47%）	3594
总计	17316	9956	27272

每个簇的光谱的颜色星等图见图 5，黄色散点为全部光谱样本分布，黑色散点为每一簇中光谱的分布。颜色星等图服从双峰分布，两端分布为红色和蓝色部分，过渡区为绿谷，可以明显看出发射线星系 cluster1 分布在蓝色区域，吸收线星系 cluster2 分布在红色区域，具有弱发射线的 cluster3 分布在绿谷，这符合早型星系大多为红色，晚型星系大多为蓝色的基本规律。

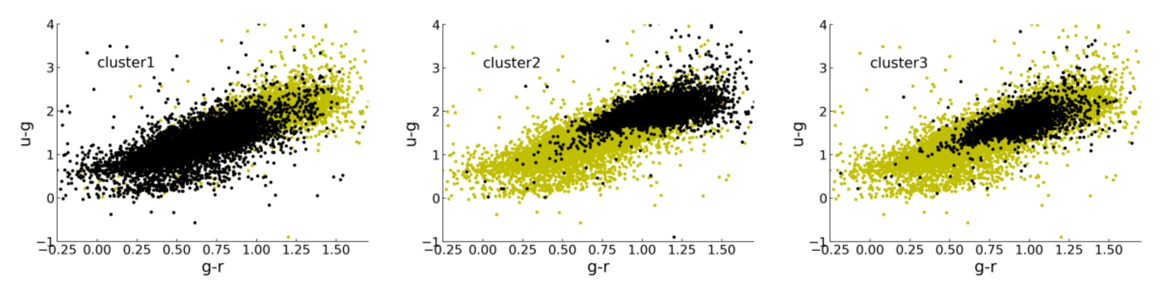


图 5 第一层聚类结果的颜色星等图。左、中、右图分别为 cluster1、cluster2 和 cluster3 的的颜色星等图，其中黄色散点为全部光谱样本，黑色散点分别为每一类光谱样本

Fig.5 Plots of *u-g* vs. *g-r* of the first layer of clustering.The plots of *u-g* vs. *g-r* of cluster1,cluster2 and cluster3 are shown on the left,middle and right.The yellow scatter points is the whole spectral samples and the black scatter points is the spectrum of each class

由实验结果可以看出 k-means 聚类算法可以快速高效地将星系光谱聚类为吸收线星系和发射线星系，对于大规模数据，k-means 聚类也能快速收敛，聚类结果能够体现出星系的物理性质，与传统分类结果基本一致，因此 k-means 聚类方法对星系分类是可行的，聚类中心可以为星系自动分类 pipeline 提供模板，与基于谱线分析得到的高信噪比模板，此模板抗噪性更强。

3.2 第二层聚类结果分析

用 CLARA 聚类将第一层聚类中的发射线星系分为 emi1-emi5 五个子类，其数目及类型见表 2，其聚类中心是类内的一条光谱（图 6 第一列）。

表 2 第二层聚类结果

Tab.2 The result of second layer clustering

类名称	数目	星系类型
emi1	2600	SF、composite、AGN
emi2	2018	SF
emi3	3576	SF、composite、AGN
emi4	1751	SF
emi5	2744	composite、AGN

与第一层聚类相同，计算每一条光谱与每个聚类中心的距离，得到第 i 个簇 cluster i 的每一个样本与第 j 个聚类中心 center j 的距离统计图，结果表明每个簇到其聚类中心最近，接近于 0，到其他聚类中心相对较远，每个簇对五个聚类中心的距离统计图都有五个明显峰值，可以表明类间区分度明显。

聚类结果与 BPT 图分类相比较，用 BPT 分类法求每一类中每条光谱的类型。BPT 图分类方法基于线强比，需要测量 $H\alpha$ 、$H\beta$ 、$[OIII]\lambda5007$ 和 $[NII]\lambda6585$ 四条谱线的线强。普遍认为星系光谱是由多种恒星光谱组合而成，首先用星族分析软件 STARLIGHT 拟合星系光谱中的恒星成分，之后用原星系光谱减去拟合谱，得到包含发射线、噪声和低频背景成分

的光谱，然后用窗口宽度为 201 的中值滤波去除低频背景成分，最后分别使用单高斯拟合来拟合 $H\beta$ 和 $[OIII]\lambda5007$ 线，用多高斯拟合来拟合 $[NII]\lambda6548$ 、 $H\alpha$ 、 $[NII]\lambda6585$ 三条谱线，利用公式（4）计算线强，其中 λ_1 和 λ_2 为谱线对应波长的两端点，$F_I(\lambda)$ 为观测流量，$F_C(\lambda)$ 为连续谱。

$$Intensity = \int_{\lambda_1}^{\lambda_2} (F_I(\lambda) - F_C(\lambda))d\lambda \quad (4)$$

由于星族成分合成过程对光谱质量要求较高和部分发射线太弱导致无法高斯拟合等问题，仅有 8122 条发射线星系光谱用 BPT 方法求得其类型，emi1-emi5 五类对应 BPT 分类结果见表 3。将每一类结果在 BPT 图中表示（图 6 中列），其中背景密度图是所有发射线星系的 BPT 图分布，红色散点是每一类中所有光谱在 BPT 图中对应的点。

表 3 第二层聚类结果与 BPT 图分类法的比较结果

Tab.3 The comparison between second layer clustering results and BPT classification method

数目(类内百分比)	SF	composite	LINER	Seyfert2	总计
emi1	789(53.86%)	406(27.71%)	60(4.09%)	210(14.33%)	1465
emi2	1297(84.00%)	177(11.46%)	55(0.97%)	15(3.56%)	1544
emi3	1587(68.38%)	485(20.90%)	71(3.06%)	178(7.67%)	2321
emi4	1386(84.31%)	127(7.73%)	8(0.49%)	123(7.48%)	1644
emi5	443(38.58%)	396(34.49%)	106(9.23%)	203(17.68%)	1148

图 6 第二列 BPT 图中，红色的经验分割线为 Kauffmann[3]等人提出的纯恒星形成星系分割线，简称 K03（公式 5），此线以下为恒星形成星系。蓝色分割线为 Kewley[4]等人提出的纯活动星系核分割线,简称 K01（公式 6），此线以上为活动星系核，混合型星系位于 K03 与 K01 分割线之间。绿色分割线为 Cid Fernandes[6]等人提出的用于区分 Seyfert2 和 LINER 的分割线,简称 CF10（公式 7），此线以上为 Seyfert2 星系，以下为 LINER 星系。

$$\log_{10}([OIII]/H\beta) = 0.61/[\log_{10}([NII]/H\alpha) - 0.05] + 1.3 \quad (5)$$

$$\log_{10}([OIII]/H\beta) = 0.61/[\log_{10}([NII]/H\alpha) - 0.47] + 1.19 \quad (6)$$

$$\log_{10}([OIII]/H\beta) = 0.01 * \log_{10}([NII]/H\alpha) + 0.48 \quad (7)$$

从聚类结果的 BPT 图和表 3 中各类星系的数量可以看出 emi1 大部分分布在 K01 分割线之下，包括恒星形成星系和混合型星系；emi2 大部分在 K03 分割线之下，有 84.00%光谱为恒星形成星系；emi3 与第一类相似，大部分为恒星形成星系，包含少量 AGN 星系；emi4 位于 K03 分割线之下，有 84.31%的光谱为恒星形成星系，不同于第二类，emi4 的 $[OIII]\lambda5007$ 与 $H\beta$ 的线强比偏大，对应聚类中心光谱，emi4 相较 emi2 发射线更强，连续谱更平缓，吸收线成分更弱；emi5 中有 61.42%的星系为复合型星系和 AGN 星系，与 emi2 和 emi4 这两类恒星形成星系相比，emi5 的聚类中心光谱的恒星成分占主导地位，发射线很弱，而 emi2 和 emi4 的聚类中心中发射线很强，占主导地位。整体来看恒星成分越少，发射线越强，星系在 BPT 图中分布越偏向于恒星形成星系，这符合恒星形成星系的特点，这类星系具有大量恒星形成区，能够观测到来自中央区域的强窄发射线，这在 emi2 和 emi4 的聚类中心光谱中也有所体现。

将聚类结果的颜色星等图画出（图 6 第三列），黄色散点是包括吸收线星系在内的所有星系光谱对应的颜色星等图，黑色散点是第二层聚类中每一类对应的颜色星等图。从 emi2

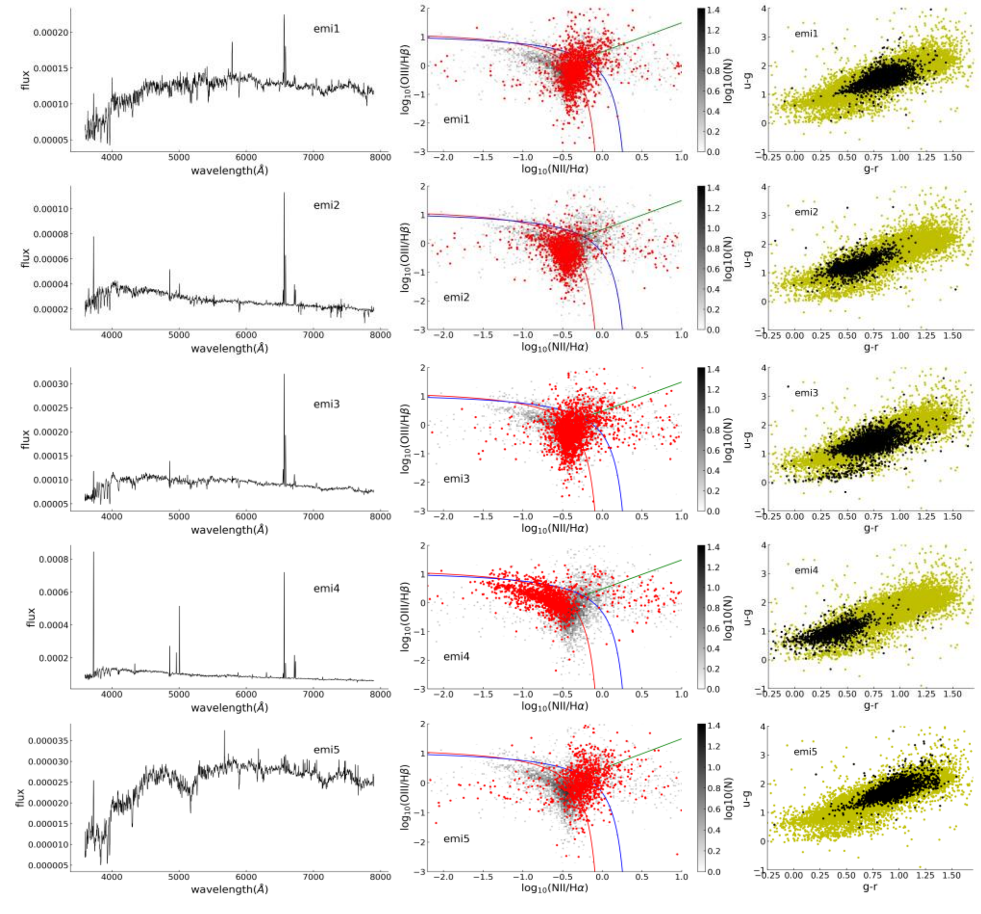


图 6 第二层聚类的聚类中心、BPT 图和颜色星等图。左、中、右列分别为聚类中心、BPT 图和颜色星等图，1-5 行分别为 emi1-emi5 类。BPT 图中黑色背景密度图为全部发射线星系样本分布，红色散点分别为每一类的光谱样本，颜色星等图中黄色散点为全部光谱样本，黑色散点分别为每一类光谱样本

Fig.6 The clustering centers , *u-g* vs. *g-r* plots and BPT diagram of the second layer of clustering.The left,middle,right column are clustering centers,BPT diagram and *u-g* vs. *g-r* plots,and lines 1-5 are emi1-emi5.In the PBT diagram,the black background density map shows the sample distribution of all emission line galaxies,and the red scatter points is sample distribution of emi1-emi5.In the *u-g* vs. *g-r* plots,the yellow scatter points is the whole spectral samples and the black scatter points is the spectrum of each class

和 emi4 可以看出 SF 更偏向于蓝色，且发射线越强颜色越蓝，emi1 和 emi3 属于绿谷，emi5 更偏向于红色，这与目前提出的 AGN 星系更可能为早型星系的观点[12]一致。同时，从 emi2、emi4 到 emi1、emi3 最后到 emi5，随着 AGN 数量的增加，在颜色星等图上反映出从蓝色到红色的变化过程，这与 Schiawinski[13]提出的 AGN 活动抑制了恒星的形成，因此它可能是星系颜色穿越绿谷的原因这一观点一致。

BPT 图分类方法步骤复杂，对光谱质量要求高，实验第二层中发射线星系能全部被 CLARA 算法划分，而 BPT 图只能分类出其中的一大部分，由此可以看出 CLARA 算法的优越性。CLARA 算法对光谱质量要求低，不需要拟合恒星成分，方法简单有效，针对大规模星系光谱能够快速有效分类，适用于大规模数据自动分析处理，同时分类结果能够很好地反映出星系的演化过程。

4.结论

针对 LAMOST DR5 星系光谱数据，使用 k-means 聚类算法成功将星系光谱分为吸收线

星系和发射线星系，与基于谱线检测的分类结果基本一致。k-means 聚类算法简单高效，适用于大规模星系光谱自动分析处理，聚类结果能够良好地反映出星系的性质，与传统分类结果基本一致，因此聚类方法对星系分类是可行的，聚类中心能够为星系光谱自动分类提供三种类型模板，相较于基于谱线分析得到的高信噪比模板，聚类中心作为模板抗噪性更强。

使用 CLARA 聚类算法将发射线星系细分类，结果与 BPT 图分类和颜色星等图分类结果存在预期的相关性，能够反映出星系的演化过程。CLARA 聚类算法对光谱质量要求较低，不需要拟合恒星成分，方法简单有效，能够直接依据谱线特征实现自动聚类，适用于大规模数据自动分析处理，能够为光谱自动分类 pipeline 提供模板。

致谢 郭守敬望远镜（大天区面积多目标光纤光谱望远镜，LAMOST）是中国科学院建设的国家重大科学项目。该项目由国家发展和改革委员会提供资金。LAMOST 由中国科学院国家天文台运营和管理。感谢审稿人提出的问题与意见，与作者进行多次深入分析与讨论，使文章内容更为严谨。

参考文献:

[1] Strateva I, Ivezic Z, G. R. Knapp, et al. Color separation of galaxy types in the sloan digital sky survey imaging data [J]. The Astronomical Journal, 2001, 122: 1861-1874

[2] J. A. Baldwin, M. M. Phillips, R. Terjevich. Classification parameters for the emission-line spectra of extragalactic object [J]. Publications of The Astronomical Society of The Pacific, 1981, 93: 5-19

[3] Kauffmann G, Heckman T M, White D M, et al. Stellar masses and star formation histories for 10^5 galaxies form the Sloan Digital Sky Survey [J]. Mon. Not. R. Astron. Soc., 2003, 341: 33-53

[4] Kewley L J, Dopita M A, Sutherland R S, et al. Theoretical modeling of starburst galaxies [J]. The Astronomical Journal, 2001, 556: 121-140

[5] Kewley L J, Groves B. Kauffmann G. The host galaxies and classification of active galactic nuclei [J]. Mon. Not. R. Astron. Soc., 2006, 372: 961-976

[6] Fernandes R C, Stasinska G, Schlickmann M S, et al. Alternative diagnostic and the ‘forgotten’ population of weak line galaxies in the SDSS [J]. Mon. Not. R. Astron. Soc., 2010, 403: 1036-1053

[7] Bolton A S, Schlegel D J, Aubourg E, et al. Spectral classification and redshift measurement for the SDSS-III baryon oscillation spectroscopic survey [J]. The Astronomical Journal, 2012, 144: 144-164

[8] Almeida J S,Aguerri J A L,Munoz-Tunon C, et al.Automatic unsupervised classification of all sloan digital sky survey data release 7 galaxy spectra [J]. The Astronomical Journal, 2010, 714: 478-504

[9] 李乡儒，胡占义，赵永恒．基于 Fisher 判别分析的有监督特征提取和星系光谱分类［J]．光谱学与光谱分析，2007, 27(9): 1891-1901

[10] 覃冬梅，胡占义，赵永恒.基于支撑向量机的天体光谱自动分类方法［J].光谱学与光谱分析, 2004, 24(4): 507-511

[11] Gauci A, Adami K Z, Abela J. Machine learning for galaxy morphology classification [J]. Mon. Not. R. Astron. Soc., 2010, 000: 1-8

[12] Mac J. Some methods for classification and analysis of multivariate observations [C].Proceedings of 5th Berkeley Symposium on Mathematics, Statistics and Probability, Berkeley, 1997: 281-296

[13] 赵国富， 曲国庆． 聚类分析中 CLARA 算法的分析与实现 ［J］. 山东理工大学学报(自然科学版)，2006(02): 45-48

[14] Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values [J]. Data Mining and Knowledge Discovery, 1998, 2(3): 283-304

Spectral Classification of Galaxies Based on Clustering Analysis

zhang xi[1,2], zhang jiannan[1], zhao yongheng[1]

(1.Key Laboratory of Optical Astronomy,National Astronomical Observatories,Chinese Academy of Sciences,Beijing 100101,China;

2.University of Chinese Academy of Sciences,Beijing 100049,China,Email:jnzhang@bao.ac.cn)

Abstract: Various large-scale sky survey plans release massive astronomical data. It is necessary to study the spectral automatic processing methods for large-scale data. It is difficult to apply the traditional galaxy spectral classification methods based on spectral line measurement or BPT diagram to automatic galaxy spectra classification pipeline directly. In contrast, machine learning method is more suitable for the classification and analysis of massive astronomical data. This paper proposes a galaxy spectral analysis method based on double hierarchical clustering. The first layer uses K-means clustering method to classify galaxy spectra into absorption line galaxies and emission line galaxies; the second layer uses Clustering Large Applications clustering algorithm to gather emission line galaxies into five subtypes. We experiment with galaxy spectral data from LAMOST DR5 and analyze the result in detail by spectral line detection, BPT diagram and color magnitude map.The experimental results show that: (1) The first layer K-means clustering can classify Galaxy spectra into absorption line galaxies and emission line galaxies successful, which are consistent with the classification results based on spectral line detection. (2) The results of CLARA cluster in the second layer can reflect different galaxy types in BPT diagram. (3) There is an expected correlation between spectral clustering results and color magnitude classification. (4) The two-layers clustering can be applied to large-scale data automatic analysis and processing. The clustering results can reflect the physical properties and evolution process of the galaxies.And the cluster centers can provide templates for automatic spectral classification pipeline.

Key words:LAMOST; Clustering; Galaxy spectra classification; Large scale spectral analyze